DOCUMENT RESUME

ED 387 522 TM 023 816

AUTHOR Mislevy, Robert J.
TITLE What Can We Learn from International Assessments?
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Kellogg Foundation, Battle Creek, Mich.; National Center for Education Statistics (ED), Washington, DC.; National Science Foundation, Washington, D.C.
REPORT NO ETS-RR-95-12
PUB DATE Mar 95
NOTE 40p.; Paper prepared for the Conference on the Use of International Educational Data (Washington, DC, February 4, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; Comparative Analysis; *Cross Cultural Studies; Educational Change; Educational Improvement; Educational Policy; *Elementary Secondary Education; *Inferences; *International Education; International Studies; *Research Utilization; Sampling

ABSTRACT

The kinds of inferences that can be drawn from international educational assessment are explored, considering the evidence that can be obtained and how it can be interpreted. International assessments have been thought of as yielding information that allows comparisons of relative achievement by country and subject, or that allows the improvement in one country from the determinants of achievement in another, or finally as a way to provide information to policymakers on the status of achievement and practices in their own countries. Issues of population definition and of sampling plans make international comparisons very difficult. The comparability of assessment tasks is complicated by the difficulty in identifying a common frame of reference. It is argued that indices of educational achievement that are to varying degrees comparable across nations can be useful, but that it must be recognized that ascertaining the relative standings of nations will tell very little about how to set educational policy or to improve instructional practice. Two tables and two figures illustrate the discussion. (Contains 48 references.) (SLD)

RESEARCH REPORT



# WHAT CAN WE LEARN FROM INTERNATIONAL ASSESSMENTS?

Robert J. Mislevy

Educational Testing Service
Princeton, New Jersey
March 1995

# What Can We Learn from International Assessments?

Robert J. Mislevy

Educational Testing Service

March, 1995

This paper was prepared for the Conference on the Use of International Educational Data, sponsored by the Board of International Comparative Studies in Education, held February 4, 1994, in Washington D.C. Support was provided by the National Center for Education Statistics, the National Science Foundation, and the Kellogg Foundation. I am grateful for discussions with Gene Johnson, Frank Jenkins, Nick Longford, Nancy Mead, Ina Mullis, Howard Wainer, Ming Mei Wang, and Kentaro Yamamoto. Figures 1 and 2 are reproduced with the permission of the College Board.

## Introduction

In broadest terms, international assessment is meant to gather information about schooling in a number of countries and somehow use it to improve students' learning. My topic is inference in international assessment. What exactly do we hope to learn from international assessments? What evidence do the kinds of information we gather provide about what we want to learn, and how do we interpret it? We will consider issues of population definition and sampling plans; of assessment exercises and background variables; of statistical advances and inferential brick walls. After a few words about the nature of evidence and inference in general, we'll take a closer look at the kinds of inferences people want to make from international assessments.

## Evidence and Inference

Inference is reasoning from what we observe and what we know to explanations, conclusions, or predictions. We are always reasoning in the presence of uncertainty. The information we work with is typically incomplete, inconclusive, and amenable to more than one explanation. We try to establish the weight and coverage of evidence in what we observe. The first question is "Evidence about what?" There is a crucial difference between *data* and *evidence*: "A datum becomes evidence in some analytic problem when its *relevance* to one or more hypotheses being considered is established. ... [E]vidence is relevant on some hypothesis if it either increases or decreases the likeliness of the hypothesis. Without hypotheses, the relevance of no datum could be established" (Schum, 1987, p. 16). Assessment data, like clues in a criminal investigation, acquire meaning only in relation to particular inferences. The same data can be direct evidence for some inferences, indirect evidence for others, and wholly irrelevant to still other inferences.

## Objectives of International Assessment

Theisen et al. (1983) distinguish three purposes that have motivated international assessment, each giving rise to conjectures with their own special inferential challenges. These purposes, and a quick summary of my conclusions about them, are listed below. Afterwards, we'll consider each in more detail.

1. *Comparing relative achievement status by country and subject.*

   Rankings of nations enjoy wide popular interest, but critics point to the difficulty of defining simple indices to compare national educational systems, and question the

utility of such comparisons even if they were technically flawless (Bracey, 1992, 1993; Rotberg, 1991). My answer to people who want comparative standings is to give them comparative standings—lots of them: in different topics, at different ages, with different kinds of tasks, both unadjusted and adjusted for factors such as national curricula and proportion of students in school. Recognizing that no single index of achievement can tell the full story and that each has its own limitations, we increase our understanding of how nations compare by increasing the breadth of our vision. Even so, however, simply ascertaining nations' relative standing tells us little about how to set educational policy or improve instructional practice.

2. *Gleaning policy implications in one nation from determinants of achievement in others.*

This objective suggests that nations' varying policies and practices constitute "natural experiments," from which we can infer the *causes* of educational achievement. But survey data cannot, by their nature, provide direct evidence about causal effects. They can, however, reveal associations that are worth investigating in studies that *can* tell us about effects of policies and instructional approaches. Technical developments allow us to better focus evidence on our conjectures about these associations in large-scale surveys, to better account for interrelationships among task performance, explanatory variables, and the hierarchical organization of schooling—but without breaking through the inferential wall to causal inference.

3. *Re-assessing in-country expenditure priorities to boost achievement.*

The idea here is to provide information to policy-makers on the status of achievement and practices within their own nations, in terms that have some international grounding. Inferences concern what students know and can do, and what is happening in classes and schools—without attempting to establish causal relationships. Even without the strong conclusions associated with experiments, nations can benefit from tracking status and change in educational conditions, and from cross-national exploratory analyses for clues about achievement. These clues can be followed up with more focused research of other, complementary, kinds, such as experiments, field studies, and in-depth analyses of specific aspects of learning. International assessments are not a source of definitive causal conclusions on which to base policy, but rather a useful component in a mix of sources of information, from which accumulating and converging results substantiate beliefs about determinants of achievement.

## Comparisons of Relative Achievement

*Comparative standings can be very useful in policy analysis, and people will construct them whether you want them to or not. It is better to compile them yourself and explain the difficulties in interpreting them rather than leaving the job to outsiders.*

Maddison, 1975, p. 170.

*To the idea that people like to have a single number we answer that usually they shouldn't get it.*

Box & Taio, 1973, p. 309.

Before one can gather evidence to rank nations, one must have an operational definition of a nation's achievement. It must determine the scope of the comparison: Is it a given learning area defined by economical multiple-choice items or by more complex but still school-like tasks, or is it broader indicators such as courses taken or school completion? An operational definition must specify the population under consideration: Is it, say, all seventeen-year-olds in a nation, just those in school, or only students instructed in the dominant language? How about nine-year olds, thirteen-year olds, and adults? Should the relationships of the tasks to the nations' diverse curricula and cultures be taken into account, and if so, how? Once these issues are decided, one must create tasks, draw samples of students, collect data, and analyze the results. Much progress has been over the past thirty years in these technical aspects of international assessments.

But different choices along each of these dimensions can lead to different rankings. Gerald Bracey (1991, 1992, 1993) and Iris Rotberg (1990, 1991) point out that the predominantly low rankings of the United States in recent assessments[1] by the International Association for the Evaluation of Educational Achievement (IEA) and the International Assessment of Educational Progress (IAEP) depend, in part, on the configurations of choices that were employed. It is fruitless to argue, however, about which choice leads to the "true" rankings. Even if comparisons of achievement are desired, educational systems and students' accomplishments vary in many aspects from nation to nation. We all know that *Consumer Reports* presents information on a wide diversity of aspects of autos.

---

[1] Especially in science and mathematics assessments (e.g., Lapointe, Mead, & Askew, 1992; Lapointe, Askew, & Mead, 1992); U.S. samples have fared better in reading.

Some, such as headroom and fuel mileage, are precisely specified in engineering terms; others, such as road feel and instrument layout, are based on expert opinion. Trade-offs are noted in overall evaluations, and direct comparisons are made only within groups of cars with similar purposes and prices, such as luxury sedans, sports cars, and minivans. Should we expect comparing nations' education systems to be so much easier than comparing automobiles? For a given level of expenditure, gathering some evidence about many aspects of education gives us a better understanding of nations' comparative status than a great deal of evidence for rankings on any single aspect.

## Scope of Comparisons

There are many aspects of achievement about which information could be gathered. Cost is always a factor in the choice; so is coverage of evidence. We will focus on what might be called "drop in from the sky" assessment, or administering common pre-specified tasks to randomly-selected students on a given day (in contrast to, for example, examining students' performance in depth and over time in the particular areas they are studying). IEA and IAEP both collect this kind of data from students. Because such comparisons, no matter how accurate or well-executed, present an incomplete picture at best, two recent reports supplemented IAEP data with such indices as rates of higher education, employment at various levels of education, and rate of engineers per 10,000 workers (Carson, Huelskamp, & Woodall's 1993 *Sandia Report;* the Salganik *et al.* 1993 OECD report). Each indicator provides information the others don't; each has limitations on its usefulness. The United States ranked first in the rate of college graduates, for example, a positive indication of commitment to education and accomplishment of students; this index does not, however, convey the capabilities of the graduates either in absolute terms or in comparison with those of other nations.

A serious limitation of any single-number index of achievement is its inability to communicate the degree and character of variation within nations. The IAEP 1991 survey of mathematics (Lapointe, Mead, & Askew, 1992), for example, ranks U.S. thirteen-year-olds near the bottom of 15 nations with a mean of 262, substantially below Korea (283) and Taiwan (285), but far above Jordan (246). Outpacing Jordan was unremarkable in light of the nations' relative stages of economic development, but placing so far behind the leaders was viewed with alarm. When these results are projected into the framework of the

1992 U.S. Trial State Assessment,[2] however, we find means for North Dakota, Iowa, and Minnesota on a par with those of Korea and Taiwan, and means for Mississippi and the District of Columbia that match Jordan's (Salganik *et al.*, 1993)![3]

## Population Definition and Sampling

The evidential value of early international assessments for international comparisons was limited by the disparity of target populations and lack of representativeness of samples. IEA's First International Mathematics Study, for example, included a survey of students in the final year of secondary study. Table 1, from Wolf (1977), shows vast differences in the proportions of students from the relevant age group enrolled in the participating nations in 1964, due to the available resources and the degree of selectivity in their school systems. Even within these diverse in-school populations, representative samples could not always be achieved because of lack of cooperation in some countries and restrictions on attainable students in others (e.g., limitation to urbanized areas or to dominant-language, hence more privileged, schools).

[[Table 1]]

Several strategies have since been employed to address these deficiencies. A first strategy has been simply to gather better evidence: defining populations more consistently across nations, designing more representative samples, and increasing cooperation rates from schools and students. A second has been to turn attention to 9- and 13-year olds, since in many countries most children of these ages are in school (compare Table 2, from Lapointe, Mead, and Askew, 1992, with Table 1). Comparability is thus improved in exchange for scope of comparison. A third strategy has been to analyze data from selected

---

[2] This projection was possible because (1) the tasks and administration conditions of the two assessments overlapped substantially and (2) they were administered to randomly equivalent groups of U.S. students at the same point in time. The degree to which results from one assessment can be linked with those of another is addressed by Linn (1993) and Mislevy (1992).

[3] And of course there exists heterogeneity within North Dakota, Iowa, and Mississippi, and within Korea, Taiwan, and Jordan as well; the point is that single-number summaries at high levels of aggregration not only obscure patterns of attainment, but that they provide little guidance for policy makers.

subpopulations of nations. Husén (1975, p. 130-131) reported that in the IEA's First International Mathematics Study (FIMS),

> *[T]he average mathematics score among United States high school graduates is far below that of all other countries. ... But when we compare the average score of the top 4 percent of the corresponding age group, a proportion selected because it represents the lowest relative number of students in any country taking mathematics, ... the range among countries is much narrower than for the entire group of terminal mathematics students. The top 4 percent of US students score at about the same level as those in other countries.*

Similarly, Westbury (1992) found that in the Second International Mathematics Study (SIMS), the top 20% of the U.S. 13-year-old students, who had taken algebra, had an average similar to top 20% of the Japanese students.

[[Table 2]]

## Context

To a participating student, an international assessment is a personal event in a social and institutional situation. Systematic differences in the ways students from different nations typically react to this situation can reduce the value of "what they do in assessment setting" as evidence about "what they know and can do" as more broadly conceived. Reports from IEA and IAEP indicate a number of threats of this kind to comparability:

- *Familiarity.* The more familiar the type of task a student is asked, the better he or she will tend to perform. The strictly-timed, no-talking-no-help, restrictive format tasks that have characterized most international assessments tend to favor students and nations in which such assessment is commonplace—just as any alternative assessment methods would favor the students familiar with them.

- *Motivation.* Archie Lapointe recalls the assembly before the administration of the 1991 IAEP assessment in Korea, to honor as "champions" of their school the students who had been selected at random to take the assessment. Might not their motivation differ from that of American 13-year-olds, excused from gym class to write impromptu essays in a survey that has no bearing on their grades? In the 1992 reading survey of the U S. National Assessment of Educational Progress (NAEP), about 3 percent of U.S. students who performed relatively well on multiple-choice tasks simply didn't bother to respond to tasks that required a more

extended written response (John Donoghue, personal communication). Motivation also depends on the relative match of tasks to students' capabilities, in content as well as format, and on the length of time students are asked to perform.

- *Style*. When students from different nations approach a set of tasks with different response styles, summary indices based on, say, proportions of correct response, can be misleading. Richard Wolf (1977, p.33) describes a problem observed in the first IEA mathematics assessment, and still encountered in the most recent IAEP:

> *Previous IEA studies had revealed interesting differences in the amount of guessing among countries. In the mathematics study, for example, students in the United States appeared to engage in a large amount of guessing; i.e., a substantial number of questions were incorrectly answered and relatively few were omitted. In contrast, Belgian students appeared to engage in very little guessing; that is, relatively few items were answered incorrectly but a substantial number of items were omitted. Belgian students, it was learned, are taught from the beginning of their school careers not to attempt to answer a question unless they are almost certain they know the answer. This is not true in the U.S., however, where guessing is often encouraged.*

The two distinct dimensions along which the performances of students from these two nations differed—how many of the tasks they attempted, and how well they did on the ones they did attempt—cannot be fully captured by any single-number summary.[4]

## Measures of Achievement

The heart of international assessment comparisons is the set of tasks on which performance is to be compared. The desire for single-number comparisons is driven by the lingering belief in universal "traits" such as "intelligence" and "mathematics achievement" which can be ascertained from performance in settings that somehow transcend culture, background, relevance, and educational experience. Now it is true that we can gather some evidence about achievement by observing performance in settings that hold some relevance across participating nations, perhaps to different degrees in varying nations; and we can, in

---

[4] Including "corrections for guessing," which produce comparable scores for students who do and do not guess *under the assumptions that they understand the adjustment and seek to maximize their score under its application.*

each nation, supplement these common tasks with additional tasks particularly relevant to that nation and perhaps selected others. But there is a big step from what we actually observe students do to what we infer about what they know, what they have learned, and what they can do in a variety of settings both in and out of school. This section considers inferential issues that concern the kinds of tasks that might be used in an international assessment, the senses in which they may or may not be comparable, and the various levels of aggregation at which one might summarize and analyze the responses.

## Comparability of assessment tasks

The very notion of quantitative comparison of nations demands a common frame of reference. The usual approach in educational measurement follows by analogy from physical measurement. One specifies the situations under which observations will be made (e.g., defining the assessment tasks and administration conditions) and the rules by which observations will be mapped to summary statements (e.g., identification of correct answers for multiple choice items, scoring guides for open-ended tasks, schemes for aggregating results from individual tasks). The final result summarizes students' *behavior* in the same way for all participants, and the precision of the summaries is reliability or measurement accuracy The usefulness of these summaries for various inferences about students' *capabilities* more broadly construed is validity.

For the most part, IEA and IAEP have relied on multiple-choice tasks to indicate achievement. The range of skills such items provoke is at best an incomplete representation of the capabilities that schooling is meant to impart. If students are familiar with tests of this type, however, the rules for what they have to do and how they will be evaluated are straightforward. Tasks that require constructed solutions probe a broader range of skills, but students' reactions to them vary even more than with multiple-choice tasks (Hambleton & Kanjee, n.d.). In addition to the motivation problems noted above, we must be concerned about the comparability of meaning across nations of both the tasks and the standards by which they are to be evaluated.

The validity of comparing students' capabilities from their performance on standard tasks erodes when the tasks are less related to the experiences of the some of the students (Estes, 1981). Suppose we ask U.S. and German students to solve statistics problems written in German. Their relative performance may be a valid indicator of their proficiency with "statistics problems written in German," but the U.S. students' performance is not a valid indicator of their understanding of statistics. The "functional equivalence" approach

to comparisons is to devise tasks for different groups which may differ on the surface, but tap comparable knowledge, skills, and strategies. The obvious fact that students in different nations (and often within a given nation) speak different languages necessitates functional rather than literal equivalence of tasks. To what extent does simply translating tasks solve the problem?

Two examples illustrate extremes in the comparability of tasks. The first is drawn from Angoff and Cook's (1988) calibration of the Scholastic Aptitude Test (SAT) and the Prueba de Aptitud Academica (PAA), college entrance examinations for high school juniors and seniors in English and Spanish respectively. Carefully translated versions of 91 multiple-choice mathematics items and 142 multiple-choice verbal items were administered to English-language SAT test-takers in the continental U.S. and to Spanish-language test takers in Puerto Rico. Figures 1 and 2 show the relative difficulties of the items in the two languages for the mathematics and verbal sections respectively. Which items are relatively easy and which are hard is quite similar in the mathematics section, but not in the verbal section. A single-number score in mathematics in either language summarizes performance in about the same way for both language groups, in the sense that which particular items a student with a given score got right would be similar regardless of whether they were in English or Spanish. In contrast, the item-by-item performances of PAA and SAT test-takers with the same overall score would generally be quite different. The mathematics scores are in this sense more comparable across languages than verbal scores. Angoff and Cook (*ibid.*, p. 6) note that

> *...these interactions were not entirely unexpected; the observation has often been made that verbal material, however well it may be translated into another language, loses many of the subtleties in the translation process. Even for mathematical items some shift in the order of item difficulties is to be expected, possibly because of differences between Puerto Rico and the United States with respect to the organization and emphasis of the mathematics curriculum in the early grades.*

[[Figures 1 & 2]]

The second example is drawn from the 1980 IEA study of written composition, which found that standard task definitions and rules of scoring did not provide a satisfactory foundation for uniformly comparable international scales of writing achievement (De Glopper, 1988, p. 74). As White and Löfqvist (1988, p. 98) explained,

> *Although it might be thought that this range of [pragmatic writing] tasks comprised a fairly basic set of writing competencies, results of the study*

> *indicated that students' familiarity with, or need for all of these uses of writing varied according to the cultural context. Thus, it was remarked that Chilean and Thai students rarely had the experience of applying for summer jobs, while in Italy the hierarchical organization of schools made it most unlikely that a student/head teacher communication of the type requested would ever occur. For Indonesian students the phenomenon of coming home to an empty house and needing to leave a written message to members of the family was unthinkable. ... Therefore, if we are to think of these writing topics as in some sense "functional" or basic, it is important to remember that the ways in which each one is focused for the purpose of the IEA study mean that for some of the students concerned, writing in these modes required a considerable imaginative leap—something quite different from the more mundane message transmission envisaged by such a range of tasks.*

We can evaluate the coverage of evidence that achievement summaries based on common assessment tasks provide with supplemental studies within nations, in which targeted, in-depth information is gathered from students who are also administered the common tasks. The resulting relationships show how performance in the common, limited-scope, assessment tasks relates to performance in students' own school settings and in real life. Depending on the learning area and the scope of the common tasks, these relationships can differ not only from nation to nation, but from school to school and from neighborhood to neighborhood within nations.

## Aggregating performance across tasks

The outcome for every individual task in an international assessment tells a story in its own right. Assessments with hundreds of tasks, like those of IEA and IAEP, tell hundreds of stories—easily too many for even a specialist to digest. For any given inference, though, certain groups of tasks, related by the skills they tap and their relation to the various curricula, tell similar stories. The fundamental law of data aggregation is that collapsing information simultaneously (1) highlights the common pattern and (2) obscures patterns that are unique. The name of the game is to determine groups of tasks that optimize evidence for inferences of interest.

Optimal levels of aggregation can differ for different inferences. As we saw in Figure 1, total scores tell most of the story for comparing the levels of performance of the PAA and SAT samples in mathematics, since the differences between groups were so

similar on all the items.[5] In this case, weighting the results differently for different items would have little effect on the summary comparison. In international assessments, however, patterns of differences among nations on tasks are often related to the degree to which the nations the emphasize topics in their curricula, as indicated by "opportunity to learn" (OTL) ratings the teachers provide (Platt, 1975, p. 46). While comparisons might be similar among tasks *within* topic groupings, comparisons can differ *across* groupings. When this happens, summaries over topics depend on the numbers of tasks that happen to be present in each topic (Wolfe, 1989). An agency contemplating comparisons based on a single-number summary should, therefore,

1. Attempt to identify interactions (some tools for doing this are discussed below);

2. Limit reports to single-number summaries only if interactions are minimal; and,

3. If interactions are found, present them and determine the stability of single-number summary comparisons with respect to alternative meaningful weightings.

If comparisons vary with importance weightings, it is the responsibility of the reporter to justify any particular choice he or she emphasizes. For example, Eugene Johnson (personal communication, 1989), finding only slight differences in the status of nations in the IAEP-88 Mathematics assessment when weighted by OTL figures from each nation in turn, offered the unweighted aggregate as a fairly representative and consensually-defined international "market-basket" of educational tasks.

Achievement indices are analogous in this respect to the Department of Labor's Consumer Price Index (CPI). The CPI is used to track changes in prices of a "market-basket" of goods and services determined through surveys as representative of a typical American. At the highest level of aggregation, the CPI itself is viewed as an (admittedly imperfect) indicator of inflation. More detailed reports present a fuller and typically more variegated picture, such as, "The CPI increased by .5% last month, but this was due

---

[5] Total score suffices for summarizing levels of observed performance on all the items, but this is not the same as saying differences in total scores are equivalent to differences in mathematical capabilities. A constant shift by which *all* items become easier or harder for members of one culture *for reasons other than the skills of interest* cannot be disentangled from this kind of data alone.

mostly to energy costs, which rose rapidly in the Northeast and moderately in the Midwest. Food prices actually fell." We note that many interest groups closely watch changes in the definition of the CPI. Government benefits such as Social Security payments are often indexed to the CPI, and a relatively minor change in the composition of the market basket can translate to hundreds of millions of dollars in a year. The Association for the Advancement of Retired People (AARP), for example, would prefer an inflation index for Social Security benefits that better matches the expenditures of retired people, including a lower proportion on housing and a higher proportion on energy.

A recent technical development concerning aggregation in international assessment merits comment. Item response theory (IRT) is a scaling approach based on the patterns of regularity among the tasks in a selected group. It characterizes *students* in terms of their overall tendency to make correct responses (for multiple-choice items) or higher-rated performances (for open-ended exercises). It characterizes *tasks* in terms of their tendency to be answered correctly or receive highly-rated responses. IRT enables comparisons on a common scale from efficient but complex designs for presenting samples of tasks to samples of students (Benefit #1). Because the task and student parameters imply probabilities of possible responses from students at any level of overall proficiency, IRT adds a layer of meaning to the achievement index (Benefit #2). IRT facilitates investigations of item-by-nation interactions, to detect differential, possibly intentional, differences across nations in tasks; this way the model can be applied just to groupings of tasks that share similar patterns across nations (Benefit #3).[6] This last point is particularly

---

[6] Sometimes we can account for the difficulties of items in terms of the skills they demand. In data from the U.S. Survey of Young Adult Literacy (Kirsch *et al,* 1993), for example, Sheehan and Mislevy (1990) were able to account for over 80-percent of the variation in document literacy IRT item difficulty parameters with descriptors of the complexity of the document and the directive and the cognitive processing requirements of the task. This opens the door to an alternative use of IRT in the upcoming international survey of adult literacy. Whether a single IRT model adequately fits document task performance across nations will be explored first. It may not, because of the interrelationships of the familiarity of content and context of documents in different cultures. But if different fits of IRT models succeed within nations and in each case the same higher-level attributes account in essentially the same way for task difficulty, then a functional equivalence among the disparate IRT scales can be achieved. Respondents could then be characterized in terms of their capabilities of carrying out, say, two-feature matching tasks in line with a

important, because IRT models don't change the fundamental law of aggregation. Model-fit investigations, such as those reported for IAEP's 1991 mathematics and science assessments (Blaise, 1992), are necessary precursors to IRT.

## Comments about International Comparisons

I stated earlier that my answer to people who want comparative standings is to give them comparative standings—lots of them: in different topics, at different ages, with different kinds of tasks; unweighted, weighted by national curricula guidelines, weighted by surveyed opportunity-to-learn; unadjusted results for the full sample, for students in selected courses of study, for students at or above selected percentiles on within-nation performance.[7] I would also provide comparisons of wholly different indices, such as the school-completion rates, school achievement data, and job-distribution-by-education characteristics. Since no single index of achievement can tell the full story and that each suffers its own limitations, we increase our understanding of how nations compare by increasing our breadth of vision—just as *Consumers Reports* informs us more fully by rating scores of attributes of automobiles. (Among, say, minivans, "best buys" score well for their cost in several categories, and strike effective balances between competing qualities such as performance and economy.) We should continue to improve the techniques we use to define indices, collect data, and analyze results for all such indices—just as *Consumers Reports* strives toward more comprehensive and more accurate measures of aspects of automobiles' safety, comfort, and performance. Only with such information can we at once compare nations on aspects we deem comparable and evaluate each in light of their own goals and resources.

While I do believe indices of educational achievement that are to varying degrees comparable across nations can be useful (for reasons discussed in the section on assessing

---

table's organization, or determining the correct entry in a three-level nested list—as evidenced in their interactions with documents reflecting the contents and contexts of their own cultures.

[7] I would even go a step further by providing multiple rankings on a given index, each differing in accordance with a randomly-selected draw of sampling-error and measurement error distributions that characterize uncertainty given a particular operational definition of achievement. Nations with barely-distinguishable values would often change ranks in these pseudo-comparisons, reducing the risk of overinterpreting small differences.

within-nation priorities), ascertaining nations' relative standing tells us little about how to set educational policy or improve instructional practice. EPA fuel mileage ratings help us to compare the cars' fuel economy, but tell engineers nothing about how to boost a given car's performance. The second motivating objective of international assessment, therefore, has been the attempt to infer the determinants of achievement.

## Determinants of Achievement

> *When studying the effectiveness of schools, investigators must ask whether the primary goal is to provide descriptive data of how things are or, rather, to estimate what outcomes would most likely occur if certain changes were introduced. If the goal is essentially description, then large-scale sample surveys provide excellent data. ... If, on the other hand, the purpose is to develop informed predictions about how educational outcomes would change in response to new or different mixes of resources, a randomized controlled field trial is preferable, almost necessary.*
>
> Platt, 1975, p. 63.

Recognizing that simple comparisons of status provide little guidance for improving education, users of assessment data, national and international alike, consistently plead for more practical advice from assessments (Viadero, 1993). What policies should we enact? How should we structure the curriculum? Which teaching practices should we follow? International assessments such as those of IEA and IAEP would appear well positioned to respond. IEA assessments, for example, solicit information from students and teachers on some 200 background variables in addition to achievement tasks, including type of school and program, student characteristics (e.g., demographics, educational background, home conditions), learning conditions (e.g., OTL, instructional practices), and kindred variables (e.g., attitudes toward education, aspirations). This section concerns the associations of background variables such as these with achievement. The central message is neither new nor optimistic: We cannot infer the causes of achievement from survey data such as those gathered in international assessment.

Does extra instruction in reading help children read better? Of course, we respond. Yet in the 1992 NAEP reading assessment, the amount of reading instruction fourth-graders receive is correlated negatively with their performance on the reading tasks (Mullis, Campbell, & Farstrup, 1993):

| | Time Spent in Reading Instruction | | |
|---|---|---|---|
| | 30-45 Minutes | 60 Minutes | 90 Minutes or More |
| Average Proficiency | 220 | 219 | 216 |

With a correlation of about -.1, reading instruction seems to *reduce* reading performance. But the average difference among students in the population who received various amounts of reading instruction—the prima facie effect—doesn't necessarily estimate the average *causal* effect of reading instruction on performance, because factors that may influence instructional time or reading performance are not taken into account in the comparison (Holland & Rubin, 1987). The NAEP report explains that the negative relationship in this example makes sense when we remember that (1) students who get extra help are usually students who seem to need extra help, and (2) students who seem to need extra help usually have low test scores to begin with.[8] The problem is that other prima facie effects we tend to interpret as causal effects if they conform to our expectations can be just as wrong for the similar reasons.

In contrast, the prima facie effect in a randomized experiment *is* an unbiased estimate of the average causal effect, because all other variables—even ones we are not aware of—are independent of the assignment of the conditions we want to compare.[9] Their effects tend to cancel out as sample size increases. Without random assignment, the

---

[8] If the correlation between NEEDING help and performance is -.6, and the correlation between NEEDING help and GETTING help is +.6, then the partial correlation between performance and receiving extra help *among students equally in need of help* would be +.4.

[9] The difference between treatment means in an experiment *is* an estimate of the average causal effect for the population involved in that experiment, and as such constitutes direct evidence for inference about the effect for that population. The same results are only indirect evidence about different populations, however, with decreasing weight as the populations differ. An experiment showing a curriculum works well in one suburban school is weaker evidence about its effect in disadvantaged urban schools than other similar suburban schools. In educational research, a case is usually built up from patterns of findings across experiments in different settings, in conjunction with survey results and studies of classroom interactions.

effects of other variables need not cancel out, so the assumptions required to infer causal effects are very strong: "[W]e might be willing to assume strong ignorability ... if each [matched group of individuals in the comparison groups] contains a very homogeneous set of individuals who tend to respond very similarly to [the alternative treatments]" (Holland & Rubin, 1987, p. 27). We might entertain a causal interpretation for the differing results between a drug with white rats in one laboratory and a different drug with the same strain of rats in a different laboratory because genetically, laboratory rats of the same strain are almost as alike as identical twins. But in international comparisons, we *don't* have this similarity of students in different nations. Even after we match for, say, age, sex, family income, and parents' education, there remain differences among students from Taiwan, the U.S., and Mozambique with respect to culture, attitude, motivation, and values that can moderate the effect of instructional practices.

This doesn't mean that studying the relationships between achievement and other variables is wholly useless. When making comparisons among nonrandomly determined groups, sometimes we can match cases or use statistical techniques such as blocking or regression to take selected variables, or covariates, into account to some extent (Rosenbaum & Rubin, 1983). Suppose we speculate that "the amount of teacher education, including preservice teacher training, is important for the performance of students, particularly in the higher grades in school" (Postelthwaite, 1975, p. 28). It may be that teachers with more training tend to teach in schools in wealthier communities, or are assigned to more advanced classes within schools. If so, the difference among students with teachers with different amounts of experience would depend on these factors as well, and, as in the reading instruction example, modify or even reverse the relationship. We would instead compare achievement among students whose teachers had different amounts of training, but who had similar scores at the beginning of the year, took comparable courses in previous years, and lived in the same kinds of neighborhoods. Other factors we don't have data to match on, or aren't even aware of, can still make the "matched cases" effect different from the causal effect, but at least we've eliminated some of the effects we knew could skew the results.

The availability of strong, well-understood covariates[10] never eliminates the potential of a large difference between prima facie and causal effects, but it does require ever stronger relationships among the studied variables and omitted variables to alter the relationship. Thus, studying associations among background variables and achievement variables is most useful as circumstantial support for conjectures about the determinants of achievement, or as a source of inspiration for new conjectures. We note in passing that while relationships with achievement are generally found for OTL and economic status variables in international surveys, findings related to more specific instructional practices have been disappointing. Thorndike (1973, p. 178) lamented the early pattern, rarely broken:

> *In general, the factors that it was possible to identify in the school are at best minimally related to reading achievement, and a relationship that is found in any country rarely appears consistently in the others. Even the variables that one might anticipate a priori would be predictors of achievement do not tend to hold up. For example, indicators of training of teachers in the teaching of reading, of size of class, and of availability of specialist teachers in the school all turn out to have either no relationship to reading achievement or a relationship the reverse of what one might anticipate.*

For pointed conjectures about the effect of particular variables, a well-designed, randomized field study with 200 students can provide stronger evidence than an international assessment with 200,000 students (Wiley & Bock, 1967).

As a reviewer of an earlier draft of this paper pointed out, however, experiments in education have not always produced the definitive results their advocates hope for. First, any experimental comparison takes place within a context. Potential explanatory variables can be eliminated as explanations of results when, by design, they are not confounded with the studied conditions, but it remains an open question as to whether the same effects would be observed in different contexts. For example, a course that works well in the suburbs may fare poorly in the inner city. Second, unbiased effects of *some* agent are estimated in an experiment, but because the difference in conditions that experimenters intended to manipulate is not always the difference in conditions that actually occurs it may be difficult to isolate the explanation. For these reasons, any experiment, like any survey,

---

[10] Which are generally lacking anyway in international assessments, as discussed in the section on Background Variables.

cannot be considered definitive: it is the accumulation of corroborating evidence in different contexts and different circumstances that ultimately sways our belief.

## An Analogy from Medical Research

Medical research illustrates how different kinds of studies complement others. The Centers for Disease Control might carry out epidemiological study to begin to learn about the cause of an epidemic. A broad range of information is gathered from the local population, concerning nutrition, lifestyle, environment, and health. It is not known which of these variables will bear any relevance to the epidemic, and there are far too many to examine them all in controlled studies. They are examined for associations among background variables and the disease. Factors such as age, sex, and complicating health conditions are taken into account as well as possible. Background variables which are still associated with a disease are so-called "risk factors." They seem to be related somehow to the disease, but we cannot conclude from this kind of study that they *cause* the disease. To learn more, laboratory experiments or clinical field trials are required.

For example, early epidemiological studies in the 1950's showed a significant correlation between smoking and lung cancer, but with few background variables taken into account. The eminent statistician Sir Ronald Fisher pointed out that this association does not prove that smoking causes cancer. "Fisher argued that smoking might only be indicative of certain genetic differences between smokers and nonsmokers and that these genetic differences could be related to the development or not of lung cancer. Fisher did feel that 'a good prima facie case had been made for further investigation'" (Holland, 1986, p. 955). He did not mean to collect greater amounts of data the same kind, with an epidemiological survey that merely asked the same questions of a bigger sample of people. This would provide additional evidence about the question, "What is the correlation between smoking and cancer?", but not about the question we are really interested in, "Does refraining from smoking reduce a person's chances of developing lung cancer?"

We cannot carry out experiments that would provide the strongest direct evidence for this conjecture, namely, assigning people at random to smoking and not-smoking conditions. Today's stronger belief that smoking causes cancer rests on converging lines of indirect evidence. First, there are survey data which take more relevant variables into account: "Among his responses to Fisher, McCurdy pointed out that lung cancer rates increase with the *amount* of smoking and that subjects who stopped smoking had lower lung cancer rates than those who did not" (Holland, 1986, p. 955). Secondly, experiments

with laboratory animals that strongly support the conjecture that smoking causes lung cancer in mice—persuasive, though still indirect, evidence about its effect on humans.

## Background Variables

If the background variables included in an international assessment are poorly defined or unreliably measured, it is difficult to ascertain their association with achievement, however defined. There are usually tradeoffs between the quality of background information and its cost, in terms of time, money, motivation, or cooperation:

> *At the end of the year, teachers completed a background questionnaire as well as the opportunity-to-learn (OTL) instrument. The OTL questionnaire required the teacher to indicate, for each of the items in the pool (180 for eighth grade and 136 for twelfth grade) whether or not the mathematics on which the item was based had been taught to the target class. The SIMS instrumentation was very demanding of time and effort of those participating. This factor undoubtedly contributed to the relatively low participation rate." [The cooperation rate of selected public districts was about 50%; of private schools, about 40%]*
>
> IEA, 1985, p. 95-96.

IEA studies have consistently found a relationship between nations' emphasis on topics and student performance on tasks in those topics (Platt, 1975). Teachers' ratings on a simple 4-level scale of the degree to which topics have been addressed is clearly an impoverished indicator of students' educational experience. Observational studies of classroom process go further, and indeed, IEA carries them out in selected assessments. These studies are expensive because they station a trained observer in the classroom for extended periods of time—probably prohibitively expensive to carry out on a large scale, given the limits of evidence from survey data about the determinants of achievement.

IEA also gathers performance data at the beginning and the end of the school year in selected assessments (IEA, 1985). This costs more than collecting data at a single point in time, but holds the promise of greater utility: Pretest performance can serve as a covariate, to allow a sharper focus on associations of achievement during the school year with OTL and reported instructional-practice variables. Matching on pretest scores (literally or statistically) approximates matching on a host of unspecified factors that were operating before the year began, such as socio-economic status and parental attitudes, although it does not control for their impact during the year.

Soliciting background data from the students themselves is quite economical, compared to ascertaining information such as home characteristics from actual observation or record searches. Especially with younger students, though, the trade-off is accuracy:

> *Some indicator systems have relied on student reports for information on background factors. ... A[n] ... analysis of the quality of responses in the High School and Beyond study provided ... sobering results. Correlation coefficients between sophomores' and parents' reports of background variables ranged from very low to quite high—for example, .21 for the presence of a specific place to study in the home; .35 for the presence of an encyclopedia in the home (an item used in the NAEP as well); .44 for mother's occupation; .50 for family income; .56 for whether the family owns or rents its residence; .81 for mother's education; and .87 for father's education (Fetters, Stowe, & Owings, 1984).*
>
> Koretz, 1992a, pp. 17-18.

In the same vein, IEA (1985, p. 23) found that "Teachers opinions about the teaching of geometry were notably at odds with their reported practices. They affirm that an intuitive approach is most meaningful, that concrete models and aides should be used and that activities to improve spatial ability should be included. But in reality the most emphasized approach was a statement of definitions." (IEA, 1985, p. 23). We could further speculate about the degree to which their reported practices are at odds with their actual practices.

## Statistical Methodology

To study the effect of teacher training on student achievement, we discussed the utility of matching students on neighborhood, courses taken, and previous scores. We would look at the difference in performance of groups of students who differ as to the training their teachers had received, but who were matched on another set of background variables we believe influence achievement. But unless we have a very large sample and few matching variables, the number of students in matched groups becomes too small to provide stable estimates. Instead of matching explicitly, we can use regression analyses to the same end, leaning on assumptions of regularity in patterns to make up for lack of data. The outcome of interest, say, a reading score, is the dependent variable. Its associations with background variables, taking into account their associations among themselves, are expressed as regression coefficients.

A regression coefficient is interpreted in the following way. Consider groups of students who were identical on all the background variables except a particular one. The

regression coefficient for that background variable is an estimate of how different the groups' performances would be as a function of how different their values are on that particular background variable. It is a way to approximate a prima facie effect in a more complicated situation. If the students had been assigned at random to levels of that background variable, the regression coefficient also estimates the average causal effect of the variable. Otherwise, as is usually the case in assessments, it does not.

In the reading example, we knew right away that the prima facie difference was not the causal effect because it didn't make sense and we could see why. It is harder to remain skeptical in regression analysis, for two reasons. The first is that the analysis is more complex than simply comparing prima facie differences, and it is easy to think that complex analyses are doing more than they really are. But regression is not more complex because it is carrying out a different kind of inference, or because the nature of the evidence is any different; it is the exactly the same kind of reasoning applied to more complex data. The second reason is that results often seem to make sense. Al Beaton (personal communication) recalls a comments he heard about the Coleman *et al.* (1966) study of educational opportunity: "Look at the regression coefficient for teacher's education. If we provided each teacher with one additional year of schooling, we'd raise students' scores by 2 points." Al replied, "But the coefficient for 'Do you have a vacuum cleaner in your home?' [a proxy for economic status] is even higher. We should just buy each kid a vacuum cleaner."

This warning about inferring causal effects is by far the most important inferential issue for regression analysis of assessment data, within and across nations. Some additional, more technical, issues should be mentioned as limitations on the potential of regression analyses of such data:

- The poor definition or unreliable measurement of background variables that erodes the strength of their relationships with achievement is reflected as artificially low regression coefficients. A partial solution that uses the same data is to use a statistical model that attempts to first estimate, then adjust for, the effects of poor measurement (more on this below). An alternative solution is to get better data in a different kind of study. Careful observation of classroom practices in a handful of classrooms provides better information about the associations between instruction and learning than does cursory self-reported data from thousands of classrooms.

- The limited range of some potential explanatory variables also precludes the possibility of finding strong relationships. Regression focuses on differences in outcomes associated with differences in background variables, so practices that are similar among schools will not show up with significant regression coefficients, even if they have large positive impacts (Wolf, 1977, p. 121). If all class sizes are similar within a nation, for example, the regression coefficient for class size will be difficult to estimate and probably not statistically significant, even if results for much larger or much smaller classes might have differed substantially.

- Because background variables are generally associated with one another (e.g., better teachers tend to be employed in schools with more resources and more economically advantaged students), there are limits to the extent to which regression analyses can disentangle their associations with performance. Finding consistent results for a given variable in a wide variety of models adds credibility to its relevance. Finding large variations or even sign changes in its coefficients under different models indicates that the data cannot support strong statements about the relationship.

- When important background factors are omitted from the survey, their effect on achievement appears in regression coefficients for related survey variables. This happened in Beaton's vacuum cleaner example. Students' opportunities to learn in their preschool years influenced their later achievement in school. Economic conditions in the home was correlated with these opportunities. Having a vacuum cleaner was correlated with economic conditions in the home. Because the Coleman study could not directly address the key variables in this chain, the positive effect of early learning on school learning showed up as a coefficient for "Do you have a vacuum cleaner?" Sound construction of assessment questionnaires and analyses of survey data must lean on results from small-scale, in-depth, field studies. These latter studies help identify variables that seem to be important in achievement, so that affordable proxies for them can be included in large-scale surveys. Information from each kind of study thus improves the next round of the other kind.

**Structural equations modeling and multi-level analysis**

Over the past decade, two extensions of regression analysis have proven useful in analysis of assessment data. They are structural equations modeling (e.g., Muthén, 1988)

and multi-level analysis (e.g., Raudenbush & Willms, 1991). By exploiting what is known a priori about the structure of associations among variables, both techniques allow us to fit models more finely attuned to the patterns within data and better focused on the conjectures of interest. Like the move from comparisons of means to multiple regression, though, neither can overcome the hurdle between model-fitting and conclusive causal inference about determinants of achievement from survey data (although the lure is seductive; sometimes structural equations modeling is even called "causal modeling").

The main idea of structural equations modeling is to build a system of regression equations that, together, incorporate hypotheses such as which variables are associated with others only because they are both associated with a third ("conditional independence"), and which observed variables are noisy measures of the same unobserved "true" value ("measurement error models"). Suppose we determine the relationship between economical but unreliable self-reported income data and actual values in a small side study. With a structural equations model, we can use this relationship to estimate the association between performance and actual income from the attenuated association between performance and self-reported income. If these hypothesized structures are consistent with the data, resulting regression coefficients are better indicators of the associations among background factors and achievement. However, the same data can generally be fit equally well with a variety of models based on different hypothetical relationships and having different implications for some conjectures. Any result that shifts or changes sign under different plausible models is not supported by the data in hand.

The main idea of hierarchical analysis is to account for the shared impact of variables in the hierarchical organization of school systems. For example, a teacher's practices affect all the students in the class, a state's funding levels affect all the schools in the state, and a nation's educational policies affect all the schools in the nation. An immediate benefit of this approach is to allow one to model the variation in achievement across subdivisions within countries, thus revealing the degree of differences among, say, states or schools within nations (recall the "surprising" results when Salganik *et al.*, 1993, showed state-by-state means with international means). Some variables that affect educational achievement operate at the level of the student, others operate at the level of the school, the district, the class, and so on. The regression coefficients for student variables in one class, for example, may be quite different from the coefficients in another class, because the instructional approaches of the teachers differ. Raudenbush and Bryk's (1988) hierarchical analysis, for example, found the relationship between SES and achievement to

be substantially more pronounced in U.S. public schools than Catholic schools. The ability to model these kinds of differences allows us to explore more facets of associations among background variables and achievement.[11] As with comparisons of means, multiple regression, and structural equations modeling, though, one cannot draw causal inferences just because a hierarchical analysis is used, and any estimates of associations that do not hold up under a variety of alternative plausible models cannot be trusted.

## Assessments of In-Country Expenditure Priorities

We have seen that international comparisons of achievement status provide little information to guide educational policy or instructional practice, and that analyses of correlates of achievement provide at best circumstantial evidence about the causes of achievement. Are there good reasons nevertheless to continue to carry out international assessments? Perhaps the best reason, in my opinion, is that policy makers need indicators of status of educational achievement and practices within their own nations. Rather than setting policy in ignorance, they can use indices to alert them of problems and successes—without, of course, identifying causes or remedies. Indices can provide incomplete but affordable data about teaching practices and learning outcomes in the many classrooms of a nation. They provide direct evidence about conjectures concerning "What are things like?" in contrast to circumstantial evidence about conjectures of "Why are things as they are?" or "What shall we do to improve them?" The role of assessment indices of achievement and practice can thus serve a function much like the CPI does for economic policy.[12]

---

[11] In addition to examining of the overall effects of variables ("achievement as outcome"), hierarchical analyses have been used to explore the association of policy variables on relationships of student background with achievement relationships within schools or classes ("slopes as outcomes") and the identification of schools in which achievement is unexpectedly high in light of student background variables (exemplary schools research, or "residuals as outcomes").

[12] Expecting educational indices from surveys to determine educational policy is like expecting economical indices to determine economic policy; the result is bad policy. A reaction to a rapidly rising CPI without deeper analysis of the underlying factors is to impose wage-and-price controls—in effect, making it against the law for the CPI to rise further!

Comparisons of automobiles benefit from having common metrics and methods for measuring certain characteristics of performance, such as standard EPA tests for fuel consumption and braking distances from specified speeds. Sometimes the comparisons are directly relevant to our purchasing decision, but even when they are not, knowing results for a wide range of vehicles increases our understanding of the values of the cars we are considering. In the same way, some of the aspects of educational achievement we need to track in our own nation are relevant to other nations as well. Each participating nation benefits from the shared expertise, techniques, and results of an international assessment. International results on achievement, for example, add a dimension of meaning for setting standards for within-nation performance. With within-nation uses in mind, Wolf (1979) stresses that sample designs for IEA assessments "must be arranged to facilitate within-country analysis. This may be more important than provision of national summaries or international comparisons."

Within the United States, NAEP and international assessments help inform the debate about the condition of educational achievement. Essentially flat profiles of NAEP trends over two decades, limited though they are in scope, serve to help evaluate the often-heard indictment that performance has declined precipitously in recent years (Bracey, 1991; Koretz, 1992b). This is not to say that achievement is satisfactory or that improvements cannot be made. Indeed, results from international assessments suggest that they can be:

> *The NAEP data add further support to our growing understanding of instructional and course-taking patterns. ... At the high-school level, large proportions of students elect to avoid mathematics courses and, to a greater extent, science courses. Even though the United States may retain a larger percentage of students in high school than many other countries, the Second International Mathematics Study found that advanced mathematics course enrollment in the U.S. was only about average. The Second International Science Study found enrollments in advanced science courses in the U.S. to be well below other industrialized nations. Despite survey findings from NAEP and other large-scale assessments, consistently revealing that students who have had more coursework also have higher achievement levels, even students in academic programs often do not enroll in advanced mathematics and science courses.*
>
> Mullis, Owen, & Phillips, 1990, pp. 61-62.

As we have also seen, however, results from the OECD report comparing nations and states shows average achievement in North Dakota and Minnesota that looks like that of the highest nations in the survey, and averages in Mississippi and the District of Columbia that look like those of developing nations. The policies and practices that

improve learning will certainly differ from one school to another. in ways we will only learn through experiments, field trials, and research into the processes of school learning.

International assessments can thus play an important role in each nation's system of educational research. Complementary ways of acquiring different forms of evidence for different purposes are all needed, ranging from the wide coverage, easier-to-get survey indicators found in international assessments to the insights of in-depth studies of the learning processes of individual students. Links among levels and kinds of research improve what we learn from each. Including measures of opportunity-to-learn from an international assessment in a more comprehensive observational study on classroom practices, as an example, adds contextual grounding and possibilities of improved analysis to the observational study.

## Conclusion

In 1973, Harvard University and IEA sponsored a conference summarized in *On Educational Policy and International Assessment* (Purves & Levine, Eds., 1975). It documented lessons researchers had learned about defining populations and achieving more representative samples, about devising and interpreting assessment tasks, and about how to analyze the data and draw justifiable inferences from it. Further progress has been made on all of these fronts since then. Nevertheless, Marshall Smith's (1975) cautionary comment on inferential issues is as timely today as it was then; the key issues Smith raises are the same ones discussed in the present paper (which was drafted before I saw Smith's commentary). Despite 30 years spent figuring out what can be learned from international assessment data and developing ways of doing it, expectations of "true" comparative rankings results and of causal conclusions from survey data remained largely unabated.

Why is this so? Perhaps Howard Gardner's conjecture about "the unschooled mind" explains it:

> *[I]n most domains of knowledge, we develop very powerful theories when we are very young. School and the disciplines are supposed to reformulate those theories and to make them more comprehensive and more accurate. As long as we stay in school, we can maintain the illusion that the effort has succeeded, but ... once we leave school, the illusion disappears and there is a five-year-old mind dying to get out and express itself. ...*
>
> *No one has to tell a kid that heavy objects fall more quickly than light objects. It's totally intuitive. It happens to be wrong. Galileo showed*

> *that it was wrong. Newton explained why it was wrong. But, like others with a robust five-year-old mind, I still believe heavier objects fall more quickly than lighter objects. ...*
>
> *The only people on whom these engravings change are experts. Experts are people who actually think about the world in more sophisticated and different kinds of ways. ... In your area of expertise, you don't think about what you do as you would when you were five years of age. But I venture to say that if I get to questioning you about something that you are not an expert in, the answers you give will be the answers you would have given before you had gone to school.*
>
> Gardner, 1993, p. 5.

A five-year-old mind thinks about international assessments in terms of games—who won and who lost?—and of prima facie differences as causal effects. This is as true today as it was true 30 years ago, quite independent of advances in the expert perspective. This paper maintains that international assessments, done well, can indeed provide useful information to help nations improve schooling—but not by becoming big enough and comprehensive enough to provide "the right answers." International assessments convey context, clues, and current conditions. But after a point, a dollar spent on international assessment to enhance this kind of information tells less for improving education than the same dollar spent to obtain different, complementary, kinds of information, as from field experiments, close observation of classroom processes, and investigations of cognitive aspects learning. It is less newsworthy and less spectacular than a definitive experiment in physics, but the gradual confluence of evidence of different kinds, of different strengths, from different sources, is the source of accumulating wisdom upon which educational policy should be grounded.

## References

Angoff, W.H., & Cook, L.L. (1988). *Equating the Scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test.* College Board Report No. 88-2/ETS RR 88-3. New York: College Entrance Examination Board/Princeton, NJ: Educational Testing Service.

Blais, J.G. (1992). *IAEP Technical Report, Vol. 2.* Princeton, NJ: The International Assessment of Educational Progress/Educational Testing Service.

Box, G.E.P., & Taio, G.C. (1973). *Bayesian inference in statistical analysis.* Reading, Mass: Addison-Wesley.

Bracey, G.W. (1991). Why can't they be like we were? *Phi Delta Kappan, 73,* 104-117.

Bracey, G.W. (1992). The second Bracey report on the condition of public education. *Phi Delta Kappan, 74,* 104-117.

Bracey, G.W. (1993). The third Bracey report on the condition of public education. *Phi Delta Kappan, 75,* 104-117.

Carson, C.C., Huelskamp, R.M., & Woodall, T.D. (1993). Perspectives on education in America. *Journal of Educational Research, 86,* 259-311.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of educational opportunity.* Washington, D.C.: U.S. Department of Health, Education, and Welfare.

De Glopper, K. (1988). The results of the international scoring sessions. In T.P.Gorman, A.C. Purves, & R.E. Degenhart (Eds.), *The IEA Study of Written Composition I: The international writing tasks and scoring scales* (pp. 59-75). Oxford: Pergamon Press.

Estes, W.K. (1981). Intelligence and learning. In M.P. Friedman , J.P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 3-23). New York: Plenum.

Fetters, W.B., Stowe, P.S., & Owings, J.A. (1984). *High School and Beyond: Quality of responses of high school students to questionairre items.* Washington, D.C.: National Center for Education Statistics.

Gardner. H. (1993). Educating the unschooled mind. Washington, D.C.: Federation of Behavioral, Psychological, and Cognitive Sciences.

Hambleton, R.K., & Kanjee, A. (n.d.). Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods. Unpublished manuscript. Amherst, MA: Department of Education, University of Massachusetts.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-960.

Holland, P.W., & Rubin, D.B. (1987). Causal inference in retrospective studies. *Research Report RR-87-7.* Princeton: Educational Testing Service.

Husén , T. (1975). Implications of the IEA findings for the philosophy of comprehensive education. In A.C. Purves & D.U. Levine (Eds.), *Educational policy and international assessment* (pp. 117-143). Berkeley, CA: McCutchen.

International Association for the Evaluation of Educational Achievement (IEA). (1985). *Second International Mathematics Study: Study report for the United States.* Champaign, IL: Author.

Kirsch, I., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America.* Princeton, NJ: Educational Testing Service.

Koretz, D. (1992a). Evaluating and validating indicators of mathematics and science education. *RAND Note No. N-2900-NSF.* Santa Monica, CA: RAND.

Koretz, D. (1992b). What happened to test scores, and why? *Educational Measurement: Issues and Practice, 11,* 7-11.

Lapointe, A.E., Mead, N.A., & Askew, J.M. (1992). *Learning mathematics* (Report No. 22-CAEP-01). Princeton: Educational Testing Service.

Lapointe, A.E., Askew, J.M., & Mead, N.A. (1992). *Learning science* (Report No. 22-CAEP-02). Princeton: Educational Testing Service.

Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.

Maddison, A. (1975). Commentary on J. Vaizey's "Implications of the IEA studies for educational planning with respect to organization and resource allocation." In A.C. Purves & D.U. Levine (Eds.), *Educational policy and international assessment* (pp. 168-175). Berkeley, CA: McCutchen.

Mislevy, R.J. (1993). *Linking educational assessments: Concepts, issues, methods, and prospects.* (foreword by R.L. Linn) Princeton, NJ: Policy Information Center, Educational Testing Service. (ERIC #: ED-353-302)

Mullis, I.V.S., Campbell, J.R., & Farstrup, A.E. (1993). *NAEP 1992 reading report card for the nation and the states.* Princeton, NJ: Educational Testing Service

Mullis, I.V.S., Owen, E.H., & Phillips, G.W. (1990). America's challenge: Accelerating academic achievement. *Report No. 19-OV-01.* Princeton, NJ: Educational Testing Service/National Assessment of Educational Progress.

Muthén, B. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer and H. Braun (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum.

Platt, W.J. (1975). Policy making and international studies in educational evaluation. In A.C. Purves & D.U. Levine (Eds.), *Educational policy and international assessment* (pp. 33-59). Berkeley, CA: McCutchen.

Postelthwaite, T.N. (1975). The surveys of the International Association for the Evaluation of Educational Achievement (IEA). In A.C. Purves & D.U. Levine (Eds.), *Educational policy and international assessment* (pp. 1-32). Berkeley, CA: McCutchen.

Purves, A.C., & Levine, D.U. (Eds.). (1975). *Educational policy and international assessment.*. Berkeley, CA: McCutchen.

Raudenbush, S.W., & Bryk, A.S. (1989). Quantitative models for estimating teacher and school effectiveness. In R.D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 205-232). San Diego: Academic Press.

Raudenbush, S.W., & Willms, J.D. (Eds.) (1991). *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective.* San Diego: Academic Press.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41-55.

Rotberg, I. (1990). I never promised you first place. *Phi Delta Kappan, 72,* 296-303.

Rotberg, I. (1991). How did all those dumb kids make all those smart bombs? *Phi Delta Kappan, 73,* 778-781.

Salganik, L.H., Phelps, R.P., Bianchi, L., Nohara, D., & Smith, T.M. (1993). *Education in states and nations: Indicators comparing the U.S. states with the OECD countries in 1988.* Report NCES 93-237. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement.

Schum, D.A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, Md.: University Press of America.

Sheehan, K.M., & Mislevy, R.J. (1990). Integrating cognitive and psychometric models in a measure of document literacy. *Journal of Educational Measurement, 27,* 255-272.

Smith, M.S. (1975). Commentary on R.L. Thorndike's "The relatiuon of school achievement to differences in the backgrounds of children." In A.C. Purves & D.U. Levine (Eds.), *Educational policy and international assessment* (pp. 111-116). Berkeley, CA: McCutchen.

Theisen, G.L., Achola, P.P.W., & Boakari, F.M. (1983). The underachievement of cross-national studies of achievement. *Compoararive Education Review, 27,* 46-68.

Thorndike, R.L. (1973). *Reading comprehension education in fifteen counties: An empirical study,* International Studies in Evaluation, Vol. III. New York: Wiley; Stockholm: Almqvist & Wiksell.

Viadero, D. (Dec. 8, 1993). NAEP Urged to Make "Report Card" More Useful. *Education Week.*

White, J., & Löfqvist, G. (1988). Pragmatic writing tasks. In T.P.Gorman, A.C. Purves, & R.E. Degenhart (Eds.), *The IEA Study of Written Composition I: The international writing tasks and scoring scales* (pp. 79-99). Oxford: Pergamon Press.

Wiley, D.E., & Bock, R.D. (1967). Quasi-experimentation in educational settings: Comment. *The School Review, 75,* 353-366.

Wolf, R. (1977). *Achievement in America.* New York: Teachers College Press.

Wolf, R. (1979). Sampling. *Bulletin 4: Secondary Study of Mathematics.* Urbana, IL: Second International Mathematics Study.

Wolfe, R.G. (1989). An indifference to differences: Problems with the IEAP-88 study. Paper presented at a research conference on the Second International Mathematics Study data, University of Illinois, Champaign, February 2, 1989.

Table 1

Percent of Age Group Enrolled in Full-Time Schooling in the Terminal Year of Secondary School in 1964*

| Country | Percent |
|---|---|
| United States | 75 |
| Belgium (Flemish) | 47 |
| Belgium (French) | 47 |
| Sweden | 45 |
| Australia | 29 |
| France | 29 |
| Hungary | 28 |
| Finland | 21 |
| England | 20 |
| Scotland | 17 |
| Chile | 16 |
| Italy | 16 |
| India | 14 |
| Netherlands | 13 |
| New Zealand | 13 |
| Thailand | 10 |
| Federal Republic of Germany | 9 |
| Iran | 9 |

* Based on Table 15 of Wolf (1977).

Table 2

Age 13 Sampling Frame for IAEP Mathematics in 1991*

| Country | % Age-Eligible in School | % In-School, Age-Eligible in Frame | Comments on Frame |
|---|---|---|---|
| Scotland | 100 | 99 | |
| United States | 100 | 98 | |
| Spain | 100 | 80 | Spanish-speaking schools; not Catalan |
| Switzerland | 100 | 76 | Public schools, 15 cantons |
| Ireland | 99.8 | 93 | |
| France | 99.7 | 98 | |
| Jordan | 98.5 | 96 | |
| Hungary | 97.8 | 99 | |
| Korea | 95.9 | 97 | |
| Israel | 95.5 | 71 | Hebrew-speaking public schools only |
| Slovenia | 95.4 | 97 | |
| Canada | 94-100 | 94 | |
| Taiwan | 90 | 100 | |
| Soviet Union | — | 60 | Russian-speaking schools, 14 republics |

* Based on Figure A.2 of Lapointe, Mead, & Askew (1992)

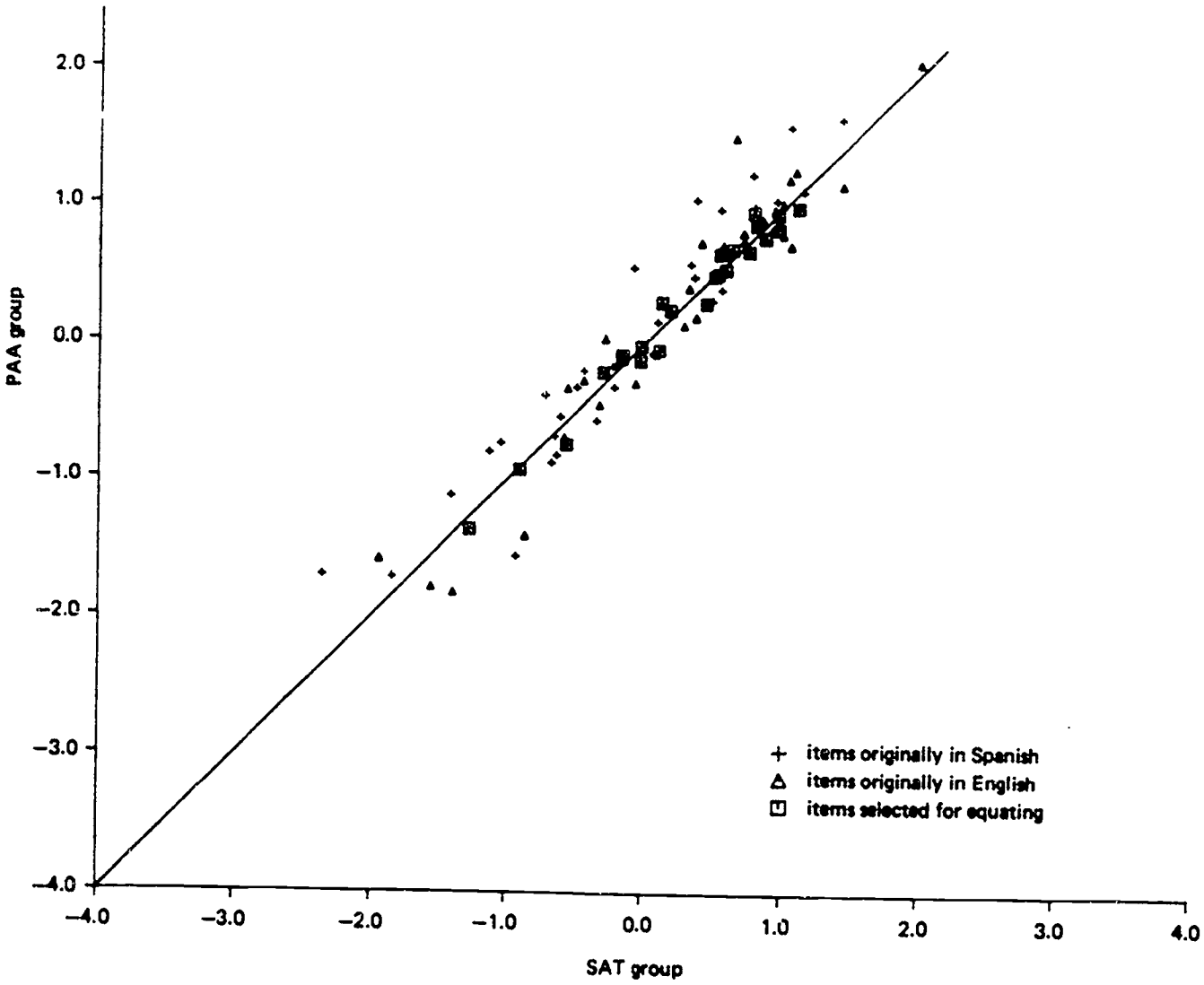


Figure 1. Plot of *b*'s for pretested mathematical items (number of items = 91).

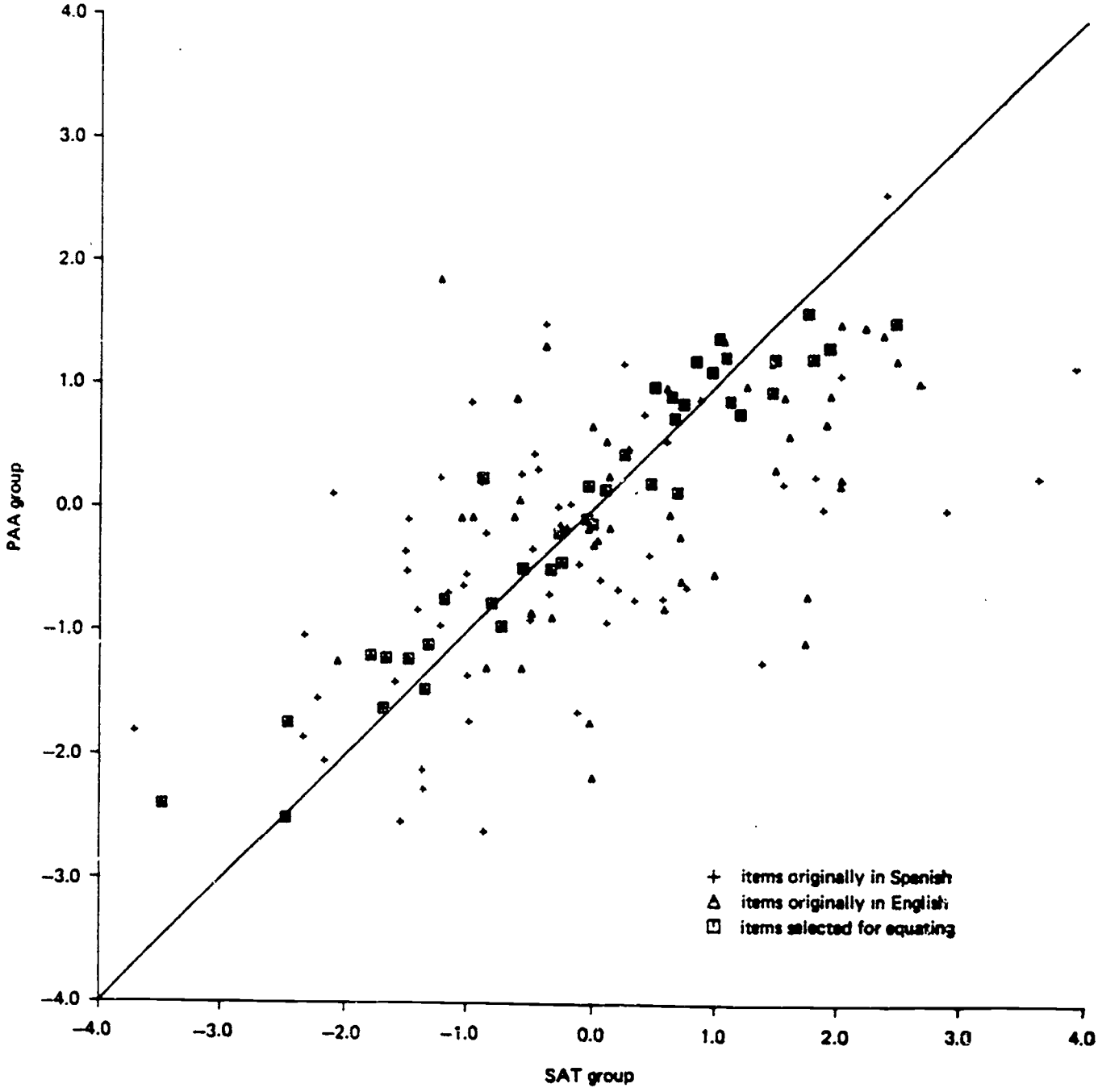


**Figure 2. Plot of *b*'s for pretested verbal items (number of items = 142).**